Sequence structures of a mouse major urinary protein gene and pseudogene compared

The EMBO Journal vol.4 no.12 pp.3159-3165, 1985

Sequence structures of a mouse major urinary protein gene and pseudogene compared

A.J.Clark[1], P.Ghazal, R.W.Bingham[2], D.Barrett[3] and J.O.Bishop

Department of Genetics, and [2]Department of Veterinary Pathology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JN, UK, and [3]Department of Biological Sciences, University of Denver, Denver, CO, USA

[1]Present address: A.F.R.C. Animal Breeding Research Organisation, West Mains Road, Edinburgh, UK

Communicated by J.O.Bishop

Laboratory mouse strains carry ~35 major urinary protein (MUP) genes per haploid genome, tightly clustered together on chromosome 4. Most belong to two main groups (Groups 1 and 2). The available evidence strongly suggests that the Group 1 genes are active while the Group 2 genes are pseudogenes. Here we present the complete sequence of a Group 1 gene and a Group 2 gene and 700 bp of flanking sequence. The sequence of the Group 1 gene is consistent with its being active. The Group 2 gene contains two stop codons and a frame-shift mutation in the reading frame defined by the Group 1 gene, and would code for a signal peptide 25 rather than 19 amino acids long. The Group 2 gene differs from the Group 1 gene in other ways: a deletion upstream of the TATA box and another in intron 3, a base change in the TATA box itself, a 2 bp duplication at the splice acceptor boundary of intron 6, an altered poly(A) addition signal and a 1-base deletion 5′ to the initiation codon. Some of these differences may explain the 10- to 20-fold higher level of Group 1 mRNA in mouse liver, and the fact that Group 1 and Group 2 transcripts are mainly spliced differently. The presence of the stop codon means that the Group 2 gene is a pseudogene in the context of the Group 1 gene. However, there is some evidence that the mature hexapeptide that it would code for may have biological activity. The 12 acceptor splice sites of the two genes all contain the identical sequence ACAG at the exon boundary. As a result this region shows an unusually high level of base-pairing homology with the splice donor site. A sequence showing a moderate to high homology with the sequence CTGAC is found between 17 and 35 bp 5′ to the acceptor site boundary in every intron.

Key words: mouse/major urinary protein/pseudogene/sequence/comparison

Introduction

The mouse major urinary proteins (MUPs) are a closely related group of small acidic proteins which are synthesised in the liver, secreted into the blood and subsequently excreted in the urine. There are ~35 MUP genes in the mouse genome (Bishop *et al.*, 1982). On the basis of nucleic acid hybridisation experiments the 35 genes can be subdivided into two groups (Group 1 and Group 2), each with ~15 members, and a small number of other genes not closely related to either group. The Group 1 and Group 2 genes are part of large units of DNA organisation which are ~45 kb long (Clark *et al.*, 1984b; Bishop *et al.*, 1985). Each unit contains one Group 1 gene and one Group 2 gene, ~15 kb apart, in a divergent transcriptional orientation (i.e., head-to-head organisation). Here we present the full sequence of the transcription units of a Group 1 and a Group 2 gene, and also some 700 bp of flanking sequence. We show that the Group 2 gene, with two stop codons and a frame-shift mutation, is a pseudogene in the context of the Group 1 gene. However, we cite evidence that raises the possibility that the hypothetical oligopeptide product of the Group 2 gene may have biological activity. Several other differences between the Group 1 and Group 2 genes were observed, some of which may impair the efficiency of transcription or translation of the latter.

Results

Figure 1A shows the basic arrangement of Group 1 and Group 2 genes and the regions of DNA sequenced. Figure 1B and C shows M13 clones that were generated, respectively, from BS6 (Group 1) and BS2,3 and sequenced. BS2,3 is the name given to a Group 2 gene which, with its flanking regions, is defined by two overlapping clones. In the case of BS6, 568 bp of 5′-flanking sequence, the 3917 bp transcription unit and 136 bp of 3′ flanking sequence were determined. Approximately 80% of the sequence was determined on both strands. The region of BS2,3 homologous to that determined for BS6 was sequenced primarily on one strand.

Determination of the Group 1 mRNA cap site

We previously described the sequence of the combined exons of BS6. The gene encodes a short mRNA of ~750 nucleotides within six exons and a long mRNA of 882 nucleotides within seven exons (Clark *et al.*, 1984a). The two forms are generated by different splicing events. The long mRNA is considerably more abundant. Previously we positioned the mRNA cap site provisionally. On the basis of two criteria, S1 nuclease protection and primer extension, we now confirm that it is located 30 ± 1 bp downstream from the TATA box (Figure 2).

Comparison of BS6 and BS2,3

Figure 3 shows the sequences of BS6 and BS2,3, aligned to maximise base-pairing homology between them. The boxes surround the exons previously defined for BS6 (Clark *et al.*, 1984a).

Insertions and deletions. The comparison shows that there are three large insertions or deletions (>17 bp) and 20 smaller insertions or deletions (<9 bp). Otherwise the two sequences are co-linear over the entire sequenced region. The most 5′ large insertion or deletion occurs within a very A-rich tract located 50 bp 5′ to the start of each transcription unit. In BS6 this tract (primarily A, occasionally interrupted by C) is 44 bp long, whereas it is only 16 bp long in BS2,3. To date, the corresponding regions of nine different MUP genes (five Group 1 and four Group 2) have been sequenced. Many show variation in the length of the A-rich tract, from a minimum of 11 bp to a maximum of 61 bp (P.Ghazal, unpublished observations). The second major interruption in the co-linearity of the two sequences occurs in the first

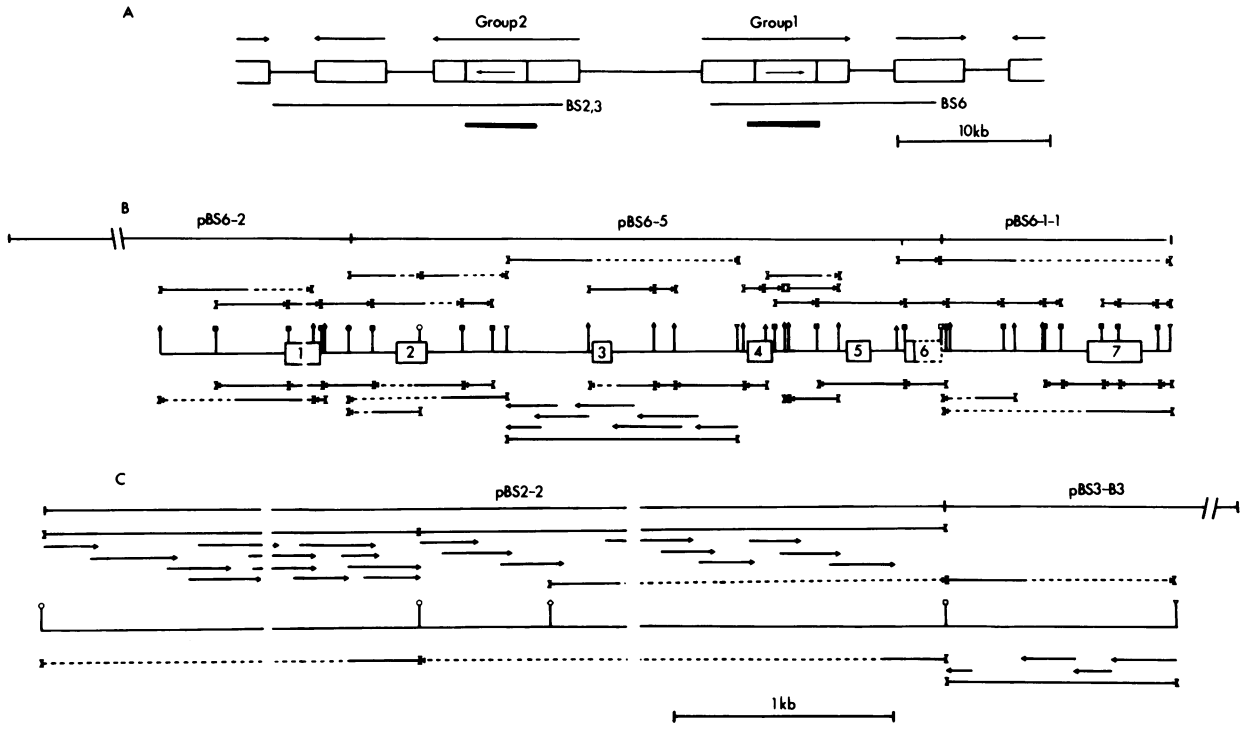

Fig. 1. Sequencing strategy for BS6 and BS2,3. **A:** The predominant arrangement of Group 1 and Group 2 genes and their flanking sequences in the BALB/c genome. Regions of inverted symmetry are shown as boxes with arrows above them. The Group 1 and Group 2 transcription units are marked as boxes containing arrows which indicate the direction of transcription. The continuous lines below show the relationship of the lambda clones to the chromosome map. BS2,3 is a composite of two Group 2 lambda clones which overlap extensively and have identical restriction enzyme sites in this region of overlap. —— Indicates the regions that were sequenced. **B:** Sequencing strategy for BS6. ⊢⊣, the plasmid subclones from which M13 clones were derived.]—[, M13 clones which were cloned at specific sites: continuous line, region sequenced; broken line, remainder of the clone which was not sequenced. →, M13 clones for which the RF was prepared and the insert progressively shortened by the method of Hong (1982). Arrows indicate the regions sequenced. Arrowheads show the direction of sequencing. The restriction map covers the region sequenced and shows the sites employed for the M13 cloning; ●, *Bam*HI; ○, *Eco*RI; □, *Hind*III; ▲, *Kpn*I; △, *Pvu*II; ▽, *Pst*I; ◇, *Aha*III; ■, *Sau*IIIA; ◆, *Alu*I. The numbered, open boxes show the positions of the exons, and the dashed extension of exon 6 shows the position of those sequences that are present in short MUP mRNA. **C:** Sequencing strategy for BS2,3. Symbols are the same as in **B**. The scale is the same for **B** and **C**.

exon within the region which codes for the signal peptide. BS6 has a 19 amino acid and BS2,3 a 25 amino acid signal peptide, the difference being due to a net insertion of six leucine residues (6 × CGT) in BS2,3. The length of this region is different in each of four Group 2 MUP genes. In contrast, the sequences of the entire signal peptide region of five Group 1 genes are identical (Ghazal *et al.*, 1985). The third major insertion or deletion is in the third intron and occurs in a region of DNA that consists primarily of runs of GT and GTT. In BS6 this region (+1537 to +1633) is 97 bp long, whereas the homologous region in BS2,3 (+1542 to +1557) is only 16 bp long. Comparable sequence data from other MUP genes are not available. However, restriction site mapping suggests that there are no large differences in length between different Group 1 genes or between different Group 2 genes.

Transcription initiation signals. The DNA sequence signals which are presumed to be required for transcription are listed in Table I. There is a possible 'CAAT' box at −109 in BS6 and −77 in BS2,3, although the sequences are considerably diverged from the published consensus, sharing only 5/9 positions, one of which is an unspecified pyrimidine in the consensus sequence. Both BS6 and BS2,3 have a consensus 'TATA' box at −31. BS2,3, however, contains a G at a position normally occupied by an A (Table I).

Splice sites. Table I also tabulates the donor and acceptor splice sites of the six introns of each gene. All 24 sites accord with the GT/AG rule and show a good agreement with the consensus sequences derived by Breathnach and Chambon (1981). In the six donor sites, BS6 and BS2,3 differ in a total of two positions (2/36 bp). Similarly the two genes differ by a total of two positions in five of the six acceptor sites (2/50 bp). The acceptor site in intron 6 of BS2,3 has a net insertion of 2 bp compared with BS6. The mRNA transcribed from Group 2 genes is mainly of the short variety which lacks exon 7 and contains an extended exon 6, while the mRNA transcribed from the Group 1 genes is mainly the longer variant which contains the short exon 6 spliced to exon 7 (Clark *et al.*, 1984a). It seems possible that the net insertion of 2 bp in BS2 may underly this difference by partially inactivating the acceptor site of intron 6.

Transcription termination signals. Most Group 1 MUP mRNA contain the 250 bp long untranslated exon 7. In this exon at +3895 there is a poly(A) addition signal (AATAAA). By comparison with the sequence of a number of MUP cDNA clones (Kuhn *et al.*, 1984; Clark *et al.*, 1985) this sequence is found to be located 22 bp 5′ to the beginning of the poly(A) tract. An identical poly(A) addition signal is present in the homologous position in the BS2,3 sequence. The less abundant short forms of Group 1 mRNA which terminate at the end of an extended exon 6 are polyadenylated at sites that relate to the rare poly(A) addition site ATTAAA at +2964 in the BS1 sequence and the usual AATAAA site at +2979 (Clark *et al.*, 1984a). The sequence in BS2,3 that corresponds in position to the first of these

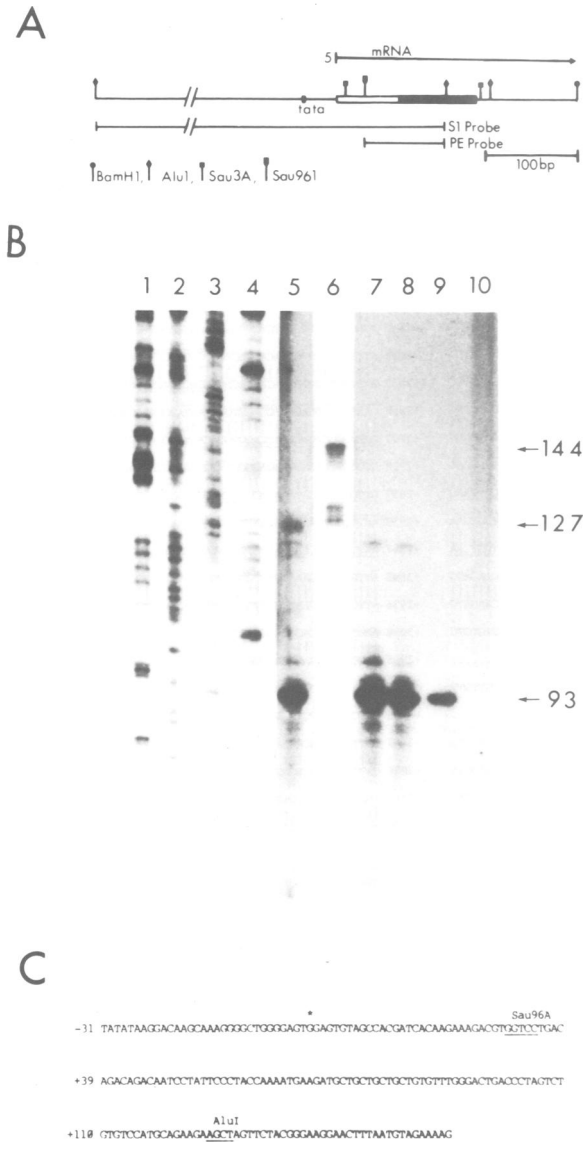

Fig. 2. S1 protection and primer extension. **A:** Restriction map of the 5′ region of MUP BS6 showing its relationship to the probes used for S1 protection and primer extension. Open and closed boxes show the untranslated and translated regions of exon 1. **B:** Electrophoretic analysis of the products of S1 protection and primer extension. **Lanes 1–4**, sequence ladder of an M13 clone used to provide mol. wt. markers. **Lane 5**, primer extension of liver poly(A)$^+$ RNA. **Lane 6**, S1 protection of liver poly(A)$^+$ RNA. **Lanes 7–9**, primer extension controls: **7**, kidney poly(A)$^+$ RNA; **8**, no RNA; **9**, primer extension probe alone. **Lane 10**, S1 protection of kidney poly(A)$^+$ RNA (S1 protection control). The primer extension probe is 93 bp long (see **C**). In the two tracks with liver poly(A)$^+$ RNA, both the S1 analysis and the primer extension analysis yield bands 127–128 bp long which positions the mRNA cap site 30 bp ± 1 bp downstream from the TATA box. An artifactual band at 144 bp is observed in **lane 6** which results from partial homology of the MUP mRNA sequences immediately 3′ to the *Alu*I sequences to the polylinker region of M13 that was present in the S1 probe. **C:** The sequence from the TATA box through the cap site (·) to beyond the *Alu*I site. The primer extension probe is the fragment from *Sau*961A to *Alu*I.

is AATAAA (+2893) and to the second GATAAG (+2907) which has not been reported to be a poly(A) addition site. There are no other AATAAA or ATTAA sequences in the region of BS2,3 within which short mRNA terminates. The present results offer a second explanation of the preponderance of short mRNA among the Group 2 transcripts: differences in the extent to which exon 7 is spliced into the mRNA may be due to differences in these 'internal' poly(A) addition signals rather than to the differences in the splice sites described above.

The coding region. The consensus sequence CCRCC has been shown to be conserved immediately 5′ to the AUG of the N-terminal methionine in a large number of eukaryote mRNAs and is proposed to be involved in ribosome binding (Kozak, 1984a). Within this consensus the R (usually A) at −3 from AUG is the most highly conserved residue, and its mutation to C in the rat pre-proinsulin gene dramatically reduced the efficiency of translation (Kozak, 1984b). The sequence immediately 5′ to ATG in BS6, CCAAA, conforms reasonably well with the consensus. In BS2,3, however, a 1 bp deletion relative to BS6 brings a C into the −3 position, thus raising a question as to whether BS2,3 transcripts would efficiently initiate translation.

Group 2 genes are transcribed much less abundantly than Group 1 genes (Clark *et al.*, 1984a). The combined exonic sequence of BS2,3 could not code for a mature MUP protein because it has stop codons in exon 1 (+156) and exon 3 (+1422) and a frame-shift mutation in exon 3 (+1472 to +1473) which generates a stop codon at +1482. In the other two frames BS2,3 contains no long open reading frames. Thus BS2,3 is a MUP pseudogene in that it has three lesions which make it untranslatable. We showed previously that three other Group 2 genes share the same stop codon in exon 1 and that the mutation therefore is probably ancestral to the Group 2 lineage (Ghazal *et al.*, 1985).

Discussion

Splice sites and intronic sequences

An interesting feature of the six acceptor sites in each of the genes is the absolute conservation of the four 3′-terminal bp. The splice acceptor site consensus sequence, derived from many genes, is NCAG, where A and G are absolutely conserved, C is present in 80% of cases, and N can be any base. In all six sites of both MUP genes this sequence is ACAG, the most notable feature being the conservation of the first A. The consensus NCAG is drawn from a large sample of different genes (Breathnach and Chambon, 1981), and would obscure such a feature of any single gene. We have therefore examined the acceptor sites of a number of genes that have multiple introns: mouse dihydrofolate reductase (Nunberg *et al.*, 1980; Crouse *et al.*, 1982; Simonsen and Levinson, 1983), alpha-fetoprotein (Law and Dugaiczyk, 1981; Eiferman *et al.*, 1981; Gerin *et al.*, 1981), alpha-amylase (Hagenbuechle *et al.*, 1981; Young *et al.*, 1981), MHC genes H-2 K-B (Weiss *et al.*, 1983) and H-2 L-D (Moore *et al.*, 1982; Evans *et al.*, 1982) and chicken alpha-2 collagen (Dickson *et al.*, 1981; Wozney *et al.*, 1981). In all cases, the terminal AG of the acceptor is absolutely conserved, but only in the case of the MUP genes is either of the two preceding nucleotides absolutely conserved.

The conserved A and C residues are complementary to the absolutely conserved G and T of the splice donor sites. We therefore asked how many base pairs would be made between five bases at the donor site of each intron (GTNNN) and the sequence NNNAC of the same intron. In nine cases three and in two cases four base pairs could be made (Table II). The probability of this arising by chance is very small (3×10^{-5}), due almost entirely to the absolute conservation of the donor site T and G residues and the acceptor site A and C residues. This highly non-random complementarity between the two regions suggests that they may come together at some stage in the splicing process. To ask

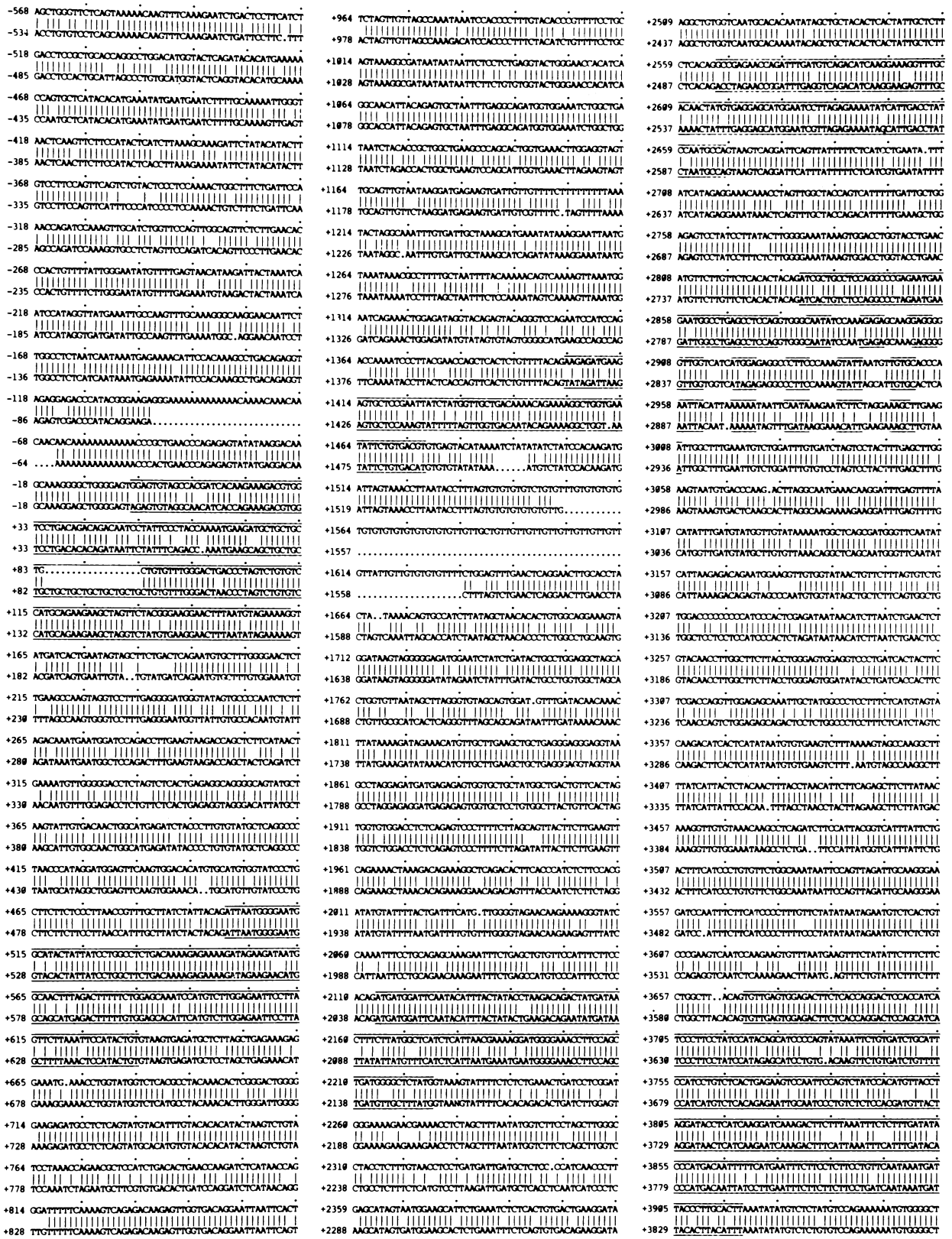

Fig. 3. Sequence comparison of BS6 and BS2,3. The sequences of BS6 (Group 1) and BS2,3 (Group 2) were aligned to maximise homology using the GAP program of Devereux and Haeberli (1984). The BS6 sequence is presented in the top line of the comparison. The regions boxed by the continuous lines show exons 1–7 of the predominant 882-bp long form of MUP mRNA (Clark *et al.*, 1984). The sequences boxed by the broken lines are those present in the shorter form of MUP mRNA. The numbers refer to the distance, in bp, from the cap site.

Table I. DNA sequence signals present in BS6 and BS2,3

Signal	Gene	Sequence	Position
Transcription initiation	BS6	GACCCATAC	−109
	BS2	GACCCATAC	−77
	Consensus*	GGYCAATCT	−80
	BS6	GAGTATATAAGG	−31
	BS2	GAGTATATGAGG	−31
	Consensus*	GNGTATAWAWNG	−30
Donor acceptor splice sites:			
Intron 1	BS6	GTATGA/TCTATTACAG	+163/+500
	BS2	GTACGA/TCTACTACAG	+180/+513
Intron 2	BS6	GTAAGT/TGTTTTACAG	+635/+1402
	BS2	GTAAGT/TGTTTTACAG	+648/+1414
Intron 3	BS6	GTGAGT/TCTTCCACAG	+1477/+2113
	BS2	GTGTGT/TCCTCCACAG	+1488/+2041
Intron 4	BS6	GTAAAG/CTTCTCACAG	+2225/+2565
	BS2	GTAANG/CTTCTCACAG	+2153/+2493
Intron 5	BS6	GTAAGT/CACACTACAG	+2668/+2830
	BS2	GTAAGT/CACACTACAG	+2596/+2760
Intron 6	BS6	GTGGGC/TGGCTTACAG	+2877/+3667
	BS2	GTGGGC/TGGCTTACACAG	+2806/+3592
	Consensus*	GTRAGT/YYYYYYXCAG	
Poly(A) addition signals:			
Exon 5	BS6	ATTAAA, AATAAA	+2964, +2979
	BS2	AATAAA, GATAAG	+2893, +2907
Exon 7	BS6	AATAAA	+3895
	BS2	AATAAA	+3819
	Consensus*	AATAAA	
Translation initiation	BS6	CCAAAATG	+67
	BS2	ACCAAATG	+66
	Consensus+	CCRCCATG	
Translation termination	BS6 (exon 6)	TGA	+2854
	BS2 (exon 1)	TGA	+156
	BS2 (exon 3)	TAA	+1422
	BS2 (exon 3)	TGA	+1482

Consensus sequences were taken from Breathnach and Chambon (1981) (*) and from Kozak (1984a) (+)

Table II. Base-pairing homology between the splice donor sites (GTNNN) and nucleotides −7 to −3 of the splice acceptor site (NNNAC) of the same intron

Intron		BS6		BS2,3	
1	NNNAC	ATTAC	3	ACTAC	3
		|||		|||	
	NNNTG	GTATG		GCATG	
2	NNNAC	TTTAC	4	TTTAC	4
		||||		||||	
	NNNTG	GAATG		GAATG	
3	NNNAC	TCCAC	3	TCCAC	3
		|||		|||	
	NNNTG	GAGTG		GTGTG	
4	NNNAC	CTCAC	3	CTCAC	3
		| ||		| ||	
	NNNTG	AAATG		NAATG	
5	NNNAC	ACTAC	3	ACTAC	3
		|||		|||	
	NNNTG	GAATG		GAATG	
6	NNNAC	CTTAC	3	TACAC	3
		| ||		|||	
	NNNTG	GGGTG		GGGTG	

Table III. Potential splice lariat junctions in the introns of MUP genes BS6 and BS2,3

Intron	BS6					BS2,3				
	Distance from junction		X	Y	Z	Distance from junction		X	Y	Z
1	21	CTTAA	3	4	3	21	CTTAA	3	4	3
2	24	CTTAC	4	5	5	24	CTTAC	4	5	5
3	22	CTGAG	4	3	3	22	CTGAG	4	3	2
4	17	CTCAC	4	4	3	17	CTCAC	4	4	
5	24	CTGAA	4	3	3	24	CTGAA	4	3	3
6	30	ATGAA	3	2	2	31	ATGAG	3	2	1

X, Y and Z, number of positions agreeing with CTGAC, CTTAC and with the complement to the splice donor site, respectively.

whether complementarity between these two regions of an intron is general, we examined the introns of the genes listed above and also those of the mouse metallothionein (Glanville *et al.*, 1981) and alpha (Mishioka and Leder, 1979) and beta (Konkel *et al.*, 1979) globin genes for evidence of complementarity between the first five bases of the donor site and the five bases before the AG of the acceptor site of the same intron. The average complementarity was 49% which, although less than the 68% found in the MUP introns, is also high. This is partly due to the absolute conservation of position 1 of the donor site and the 80% occupancy of position −3 of the acceptor site by C, but also to the fact that donor site positions 3−5 are nearly always purines while positions −5 to −7 of the acceptor site are nearly always pyrimidines. Thus elevated complementarity between the two regions is very common. If they do associate during splicing, this could follow the association of the donor (Mount *et al.*, 1983; Kramer *et al.*, 1984) and possibly also the acceptor sites (Lerner *et al.*, 1980; Rogers and Wall, 1980) with U1 snRNP, but would presumably precede the formation of the G5′-2′A lariat junction 20 or so bases upstream (Ruskin *et al.*, 1984).

Keller and Noon (1984) discovered the consensus CTGAC 20−55 nucleotides from the acceptor site boundary in a number of introns. During the search, the A residue was required to be present because in some cases it is known to participate in the junction point of the lariat splicing intermediate (Ruskin *et al.*, 1984). It was suggested that during splicing a transient base-pairing interaction occurs between this site and the splice donor site. We searched the MUP gene introns for three pentamer sequences, CTGAC itself, CTTAC which is the complement of the donor splice consensus, and the complement of the donor splice site of the intron under scrutiny. The most consistent results were obtained with CTGAC. In every intron, between nucleotides 17 and 35, there is a sequence that matches CTGAC in either four (eight cases) or three (four cases) positions (Table III). Overall, the match of these sites to CTGAC (73%) is greater than to CTTAC (69%) or to the different donor sites of the separate introns (60%). Given the selection of the A residue, we would expect this degree of matching, or better, to occur in random DNA once per 46 bases. We observe it once per 19 bases, which is not strikingly more frequent. It seems likely, nevertheless, that this technique identifies the A residue at the lariat junction in most if not all cases.

Group 2 genes are pseudogenes in the context of Group 1 genes

While the available evidence indicates that the Group 1 genes are true genes (see Clark *et al.*, 1985), all of the Group 2 genes so far examined are putative pseudogenes. BS2,3 carries three lesions in its protein coding sequence and could not be translated

to yield a protein with the mol. wt. of MUP. Partial sequence analysis of three other Group 2 genes has shown that they all contain the same stop codon in exon 1 (Ghazal *et al.*, 1985). It is likely that all Group 2 genes in the BALB/c genome share this lesion, and are descended from the same ancestral gene. Other sequence differences between BS6 and BS2,3, most of which might affect transcription or translation, are (i) upstream and intronic deletions that may affect enhancer function, (ii) a substitution of G for A in the TATA box region that may affect the strength of the promoter, (iii) a small duplication in the splice-acceptor site of intron 6 that may favour the formation of the shorter form of mRNA, (iv) an alteration in one of the poly(A) addition signals of short form mRNA (ATTAAA→AATAAA) that also may favour the formation of the shorter mRNA, (v) an alteration in the translation initiation signal CCAAA→ACCAA that may impair the efficiency of translation and (vi) an in-frame increase in the length of the signal peptide region.

A possible function for the truncated product of the Group 2 gene

Some Group 2 genes are probably transcribed to yield a short mRNA (Clark *et al.*, 1984a) although the steady-state mRNA level is much less than that observed for Group 1 genes (<10%). If the Group 2 transcripts are translated and if the polypeptides are then processed, the products will be peptides six amino acids long with a mol. wt. of 630. Such small peptides would be rapidly excreted into the urine.

Mouse urine contains androgen regulated agents that dramatically accelerate the onset of puberty when administered to young females (Vandenbergh *et al.*, 1975). One is probably a protein, with a mol. wt. >12 000, i.e., consistent with the mol. wt. of MUP. The activity of this agent largely survives proteolysis, but becomes dialysable. The second agent has a mol. wt. of 860, and seems to be one or more of a mixture of oligopeptides (Vandenberg *et al.*, 1976). These apparently contradictory observations can be reconciled by a hypothesis based on the structure of the MUP genes. We suggest that the protein agent is MUP, the active part of the molecule being the six N-terminal amino acids, and that the dialysable agent is the hexapeptide coded for by the Group 2 genes. Proteolysis of the protein agent would release dialysable fragments containing the N-terminal hexapeptide. The sequences of the two hexapeptides are quite similar: Group 1, N-Glu-Glu-Ala-Ser-Ser-Thr; Group 2, N-Glu-Glu-Ala-Arg-Ser-Met.

Group 1 and Group 2 genes have randomly diverged

BS6 and BS2,3 are members of the two major groups of MUP genes in the BALB/c genome. The numbers of Group 1 and Group 2 genes are approximately equal (Bishop *et al.*, 1982). This is because the predominant organisation of the MUP locus is an array of 45 kb domains each containing a Group 1 and a Group 2 gene linked in a divergent orientation (Clark *et al.*, 1984b; Bishop *et al.*, 1985). We have presented the sequence of BS6 and BS2,3 over a homologous region ~4.5 kb in length that includes the entire transcription unit as well as 5′ and 3′ flanking sequences. The most obvious differences between the two sequences are the three long insertions/deletions. In each case these occur in regions of 'simple sequence' DNA suggesting that they may have been created by 'slippage' during DNA synthesis or repair (Ghosal and Saedler, 1978). In general, the divergence between the two sequences is uniformly spread across the region sequenced (Table IV). Thus no recent gene correction has occurred between the two genes such as has been observed between human $^{G}\gamma$ and $^{A}\gamma$ globin genes (Slightom *et al.*, 1980). The exons and introns of BS6 and BS2,3 have diverged to about the same extent. Comparisons between other active genes indicate that, in general, intronic sequences diverge more rapidly than exonic sequences (Perler *et al.*, 1980; Efstratiadis *et al.*, 1980) presumably because introns have lesser selective constraints acting on them. That this is not the case in the comparison of the two MUP genes possibly indicates that the ancestral BS2,3 pseudogene was free to diverge at the same rate in both introns and exons. Group 2 genes, however, are reasonably well conserved amongst themselves and we have drawn from this observation the conclusion that the 45-kb domain, rather than the individual MUP gene, is the unit of evolutionary change of the majority of MUP genes (Clark *et al.*, 1984b; Bishop *et al.*, 1985).

Table IV. Divergence between BS6 and BS2,3

Region	Divergence (%)
Full length	13.4
5′ flanking region	11.1
Transcription unit	12.6
mRNA	13.1
Translated mRNA	11.5
Non-translated mRNA	15.5
Intronic sequences	13.1
3′ flanking region	15.6

The divergence between the two genes was estimated over the regions indicated. In this analysis each base change was scored as 1, as was each insertion/deletion, irrespective of size.

Materials and methods

Cloned DNA

The isolation of MUP genomic clones and subclones is described in Clark *et al.* (1982, 1984b) and Bishop *et al.* (1982). The propagation of bacteriophage and plasmid clones and the isolation of DNA were carried out as described (Clissold and Bishop, 1982; Clark *et al.*, 1982; Bishop *et al.*, 1982).

DNA sequencing

To obtain the complete 4 kb sequences of BS6 and BS2,3, fragments of plasmid pBS6-2, pBS6-5, pBS6-1-1, pBS2-2 and pBS3B-3 were cloned into M13mp7, 8 or 9 and sequenced by the dideoxy nucleotide method, essentially as described by Sanger *et al.* (1977) and Anderson *et al.* (1980). Two main strategies were employed to ensure that continuous stretches of sequence would be generated. (i) The cloned fragments were digested with restriction enzymes that cleave 4 bp recognition sites and 'shotgunned' into M13 vectors. (ii) Larger subfragments were cloned into M13mp8, replicative forms were prepared and a second generation of M13 clones containing progressively shorter fragments was isolated by the method of Hong (1982).

S1 nuclease protection

The probe was a 696 bp *Alu*I fragment, extending from nucleotide +127 to nucleotide −568 in the BS6 sequence (Figure 3), and cloned at the *Hin*cII site of M13mp7. The single-stranded M13 clone was annealed to the sequencing primer and the strand complementary to MUP mRNA was uniformly labelled using the Klenow fragment of DNA polymerase I (Boehringer). The double-stranded region thus created was digested with *Eco*RI and the fragment lying between the two *Eco*RI cloning sites of the vector (sp. act. $10^7 - 10^8$ c.p.m./μg) was purified on a 5% polyacrylamide gel. An aliquot of the probe (20 000 c.p.m.) was co-precipitated with 1 μg of total poly(A)$^+$ RNA and redissolved in 10 μl of 40 mM Pipes (pH 6.4), 1 mM EDTA, 0.4 M NaCl, 80% formamide. Samples were denatured at 85°C for 15 min and incubated at 50°C for 4 h. Samples were digested with 250 U/ml S1 nuclease (Sigma) at 37°C for 1 h in 100 μl of 0.28 M NaCl, 0.05 M NaAc (pH 4.6), 4.5 mM $ZnCl_2$ and 10 μg/ml single-stranded salmon sperm DNA, phenol extracted, precipitated twice with ethanol and resuspended in 3 μl of formamide dye mix.

Primer extension (from Ghosh et al., 1981)

The primer extension probe was the 93 bp *Alu*I-*Sau*961 fragment between nucleotides +34 and +127 in the BS6 sequence (Figure 3). This was prepared and annealed to poly(A)$^+$ RNA in essentially the same way as the S1 protection probe (above). Annealing was terminated by the addition of 250 μl ice-cold

0.3 M NaAc (pH 7.0) followed by two ethanol precipitations. The pellet was resuspended in 50 μl 50 mM Tris-HCl (pH 8.3), 6 mM $MgCl_2$, 40 mM HCl, 10 mM DTT with 1 mM of each deoxynucleotide triphosphate and 1 unit of AMV reverse transcriptase was added. After equilibration on ice for 5–10 min, the reaction mixture was incubated for 3 h at 37°C. NaOH was then added to 0.2 N and the incubation continued for a further 1 h. The reaction mixture was neutralised with 10 N HCl, phenol extracted, and ethanol precipitated twice. Pellets were resuspended in 3 μl of formamide dye mix and loaded on 6% sequencing gels.

Acknowledgements

We thank Morag Robertson and Melville Richardson for technical assistance, and the MRC and Cancer Research Campaign for financial support.

References

Anderson,S., Gait,M.J., Mayol,L. and Young,I. (1980) *Nucleic Acids Res.*, **8**, 1731-1743.
Bishop,J.O., Clark,A.J. Clissold,P.M., Hainey,S. and Francke,U. (1982) *EMBO J.*, **1**, 615-620.
Bishop,J.O., Selman,G.G., Hickman,J., Black,L., Saunders,R.D.P. and Clark, A.J. (1985) *Mol. Cell. Biol.*, **5**, 1591-1600.
Breathnach,R. and Chambon,P. (1981) *Annu. Rev. Biochem.*, **50**, 349-383.
Clark,A.J., Clissold,P.M. and Bishop,J.O. (1982) *Gene*, **18**, 221-230.
Clark,A.J., Clissold,P.M., Al-Shawi,R., Beattie,P. and Bishop,J.O. (1984a) *EMBO J.*, **3**, 1045-1052.
Clark,A.J., Hickman,J. and Bishop,J.O. (1984b) *EMBO J.*, **3**, 2055-2064.
Clark,A.J., Chave-Cox,A., Ma,X. and Bishop,J.O. (1985) *EMBO J.*, **4**, 3167-3171.
Clissold,P.M. and Bishop,J.O. (1982) *Gene*, **18**, 211-220.
Crouse,G.F., Simonsen,C.C., McEwan,R.N. and Schimke,R.T. (1982) *J. Biol. Chem.*, **256**, 8407-8415.
Devereux,J. and Haeberli,P. (1984) Program Library of the University of Wisconsin Genetics Computer Group.
Dickson,L.A., Ninomya,Y., Bernard,M.P., Pesciotta,D.M., Parsons,J., Green,G., Eikenberry,E.F., de Crombrugghe,B., Vogeli,G., Pastan,I., Fietzek,P.P. and Olsen,B.R. (1981) *J. Biol. Chem.*, **256**, 8407-8415.
Efstratiadis,A., Posakony,J.W., Maniatis,T., Lawn,R.M., O'Connell,C., Spritz, R.A., De Rich,J.K., Forget,G.G., Weissmann,S.M., Slightom,J.L., Blechl, A.E., Smithies,O., Baralle,F.E., Shoulders,C.S. and Proudfoot,N.J. (1980) *Cell*, **21**, 653-668.
Eiferman,F.E., Young,P.R., Scott,R.W. and Tilgham,S.M. (1981) *Nature*, **294**, 713-718.
Evans,G.A., Margulies,D.H., Camerini-Otero,R.D. and Seidman,J.G. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 1994-1998.
Gerin,M.B., Cooper,D.L., Eiferman,F., Van de Rijn,P. and Tilghman,S.M. (1981) *J. Biol. Chem.*, **256**, 1954-1959.
Ghazal,P., Clark,A.J. and Bishop,J.O. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 4182-4185.
Ghosal,D. and Saedler,H. (1978) *Nature*, **275**, 611-617.
Ghosh,T.K. Reddy,V.D., Taitak,M., Lebowitz,P. and Weissmann,S.M. (1981) *Methods Enzymol.*, **65**, 580-594.
Glanville,N., Durnam,D.M. and Palmiter,R.D. (1981) *Nature*, **292**, 267-269.
Hagenbuechle,O., Tosi,M., Schibler,U., Bovey,R., Wellauer,P.K. and Young, R.A. (1981) *Nature*, **289**, 643-646.
Hong,G.F. (1982) *J. Mol. Biol.*, **158**, 539-549.
Keller,E.B. and Noon,W.A. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 7417-7420.
Konkel,D.A., Maizel,J.V. and Leder,P. (1979) *Cell*, **18**, 865-873.
Kozak,M. (1984a) *Nucleic Acids Res.*, **12**, 857-872.
Kozak,M. (1984b) *Nature*, **308**, 241-246.
Kramer,A., Keller,W., Appel,B. and Luehrmann,R. (1984) *Cell*, **38**, 299-307.
Kuhn,N.J., Woodworth-Gutai,M., Gross,K.W. and Held,W.A. (1984) *Nucleic Acids Res.*, **12**, 6073-6090.
Law,S.W. and Dugaiczyk,A. (1981) *Nature*, **291**, 202-205.
Lerner,M.R., Boyle,J.A., Mount,S.M., Wolin,S.L. and Steitz,J.A. (1980) *Nature*, **283**, 220-224.
Mishioka,Y. and Leder,P. (1979) *Cell*, **18**, 875-882.
Moore,K.W., Sher,B.T., Sun,Y.H., Eakle,K.A. and Hood,L. (1982) *Science (Wash.)*, **215**, 679-682.
Mount,S.M., Pettersson,I., Hinterberger,M., Karmas,A. and Steitz,J.A. (1983) *Cell*, **33**, 509-518.
Nunberg,J.H., Kaufman,R.J., Chang,A.C.Y., Cohen,S.N. and Schimke,R.T. (1980) *Cell*, **19**, 355-364.
Perler,F., Efstratiadis,A., Lomedico,P., Gilbert,W., Kolodner,R. and Dodgson, J. (1980) *Cell*, **20**, 555-565.
Rogers,J. and Wall,R. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 1877-1879.
Ruskin,B., Krainer,A.R., Maniatis,T. and Green,M.R. (1984) *Cell*, **37**, 415-427.
Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463-5467.
Simonsen,C.C. and Levinson,A.D. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 2495-2499.
Slightom,J.L., Blechl,A.E. and Smithies,O. (1980) *Cell*, **21**, 627-638.
Vandenbergh,J.G., Whitsett,J.M. and Lombardi,J.R. (1975) *J. Reprod. Fertil.*, **43**, 515-523.
Vandenbergh,J.G., Finlayson,J.S., Dobrogosz,W.J., Dills,S.S. and Kost,T.A. (1976) *Biol. Reprod.*, **15**, 260-265.
Weiss,E., Golden,L., Zakut,R., Mellor,A., Fahrner,K., Kvist,S. and Flavell,R.A. (1983) *EMBO J.*, **2**, 453-462.
Wozney,J., Hanahan,D., Boedtker,H. and Doty,P. (1981) *Nature* **294**, 129-135.
Young,R.A., Hagenbuechle,O. and Schibler,U. (1981) *Cell*, **23**, 451-458.

Received on 27 June 1985; revised on 16 September 1985